The Journal of Computational Science and Engineering.
ISSN: 2583-9055

Spatio-Temporal Attention Mechanisms in Video-Based Visual Question Answering: A Comprehensive Review

Dunga SaiVenkat Pranav Ritvik[1], A.S. Venkata Praneel[2], P Ramaiah Chowdary[3]

[1,2,] Department of Computer Science and Engineering,
GST, GITAM [Deemed to be University] Visakhapatnam AP India.
[3]Department of Information Technology,
Sir C R Reddy college of Engineering, Eluru, AP India.
dsvpritvik@gmail.com[1], praneelsri@gmail.com[2] , ram.tech111@gmail.com[3]

Keyword:
VideoQA, VQA, Dual-LSTM Spatio-Temporal Attention, K-PSTANet, TS-STMAC.

ABSTRACT

Video Question Answering (VideoQA) has become an essential job in computer vision and natural language processing interfaces, requiring models incorporating spatio-temporal reasoning with semantic understanding. This review synthesizes advancements in methodologies designed to tackle VideoQA challenges, with a focus on three state-of-the-art approaches: Dual-LSTM Spatio-Temporal Attention, Knowledge-Based Progressive Spatial-Temporal Attention Network (K-PSTANet), and Two-Stream Spatiotemporal MAC Network (TS-STMAC). These methods employ unique attention processes, multi-step reasoning, and external knowledge to solve varied datasets and problems. By comparing their performance across benchmarks, including YouTube-QA, TGIF-QA, MSVD-QA, and ActivityNet-QA, we highlight significant advances and areas for further improvement. This paper comprehensively analyzes the growth of VideoQA models and their promise to increase real-world video understanding.

Corresponding Author: praneelsri@gmail.com

INTRODUCTION

Video Question Answering (VideoQA) [1,2,3] stands as a significant breakthrough at the interface of computer vision and natural language processing, bringing together two of the most exciting fields in artificial intelligence. Unlike classic image-based Visual Question Answering (VQA), which functions within the limitations of static images, VideoQA extends the challenge to the domain of movies [4], encompassing both spatial and temporal aspects of data. This introduces complexities on multiple layers wherein the computers need to browse through sequences of frames to capture motion dynamics and reason about relationships between entities over time. VideoQA enables machines to answer natural language queries about video content. This capacity is transformative for applications such as automatic video summarization, real-time surveillance systems, autonomous robots, multimedia retrieval, and interactive educational aids.

At the heart of VideoQA is spatiotemporal attention, a process that permits AI models to focus selectively on the most relevant areas of a video. Spatial attention [5,6,7] allows models to emphasize critical regions inside individual frames, such as objects, human behaviors, or unique visual cues. In contrast, temporal attention [8,9,10] guarantees that the model responds to key

events across a period of frames. Balancing these two competing elements, spatiotemporal attention focuses on the inherent weaknesses of films, namely the effects of redundancy across frames, occlusions, and dynamic object interactions.

For example, dual architectures such as dual-LSTM [11] introduce temporal memory that captures useful visual and textual information at each frame. Meanwhile, multi-step reasoning frameworks progressively refine attention over video information, allowing the model to align visual cues with a question better. The methods significantly improve the capability of identifying complex interactions, movement patterns, and causal links inside the movies. While the promise of VideoQA and spatiotemporal attention is tremendous, numerous obstacles remain. First, correctly combining motion signals with static frame analysis requires sophisticated models capable of reasoning across multiple timescales. Second, detecting and contextualizing keyframes in lengthy or noisy video data demands robust preprocessing and fast model designs. Lastly, redundancy in video sequences poses a unique challenge since the person is trying to maintain context while filtering the unnecessary frames.

Datasets like TGIF-QA, MSVD-QA, and ActivityNet-QA motivated research by creating benchmarks for spatiotemporal reasoning. However, problems concerning generalization and domain adaptability still need to be addressed with less application towards real-world scenarios. Furthermore, as video content grows more diverse and complicated, the demand for models that combine multimodal information—combining audio, textual metadata, and video—will expand.

This review study aims to elaborate on all state-of-the-art developments in VideoQA and spatiotemporal attention procedures. We are introducing new research regarding critical methodologies, key datasets, and specific crucial bottlenecks the field is currently facing. We are offering avenues for further inquiry within the frameworks of integrating knowledge [12,13] from sources external to the model, multimodal reasoning, and more interpretable designs for models. Thus, by filling the gap between vision and language for real-world scenarios, this VideoQA research may hold promise in constructing intelligent systems that will intelligently interact smoothly in dynamically changing environments.

DATASETS

The progression of VideoQA has been underpinned by carefully curated datasets. These benchmarks provide the foundation for model training and establish evaluation standards. Below, we detail the significant datasets that have shaped the field.

1.1 TGIF-QA

The TGIF-QA [14] dataset extends the scope of VQA to video-based issues. This dataset contains 103,919 QA pairs extracted from 56,720 animated GIFs designed to examine spatiotemporal reasoning. This dataset has four task types used:

Repetition Counting [15]: This counts the repetitions of a particular action.

Repeating action: This job is characterized as a multiple-choice question, identifying an activity repeated in a particular video.

State Transition Identification [16]: This task is also majorly used in videos that ask about transitions of states, including facial emotions (e.g., from pleased to sad), actions (e.g., from jogging to running), places (e.g., from the table to the shelf), and item attributes (e.g., from complete to empty).

Frame QA: The fact is that questions in this work can be answered from any one of the frames in a video. However, it completely depends on the video content; it can be any frame in a particular video or one particular form of a video. This dataset bridges the gap between static image tasks and the temporal reasoning necessary for video content.

1.2 MSVD-QA and MSRVTT-QA

It uses tasks created based on open-ended video content, such as short segments or lessons. MSVD-QA [17] Based on Microsoft Video Description Corpus, more than 50,000 QA pairs. MSRVTT-QA [17]. It incorporates QA pairs of MSR-VTT corpus based on its area of action, object, and scene-based knowledge. Both datasets test a model's capacity to reason about short, descriptive clips and diverse question types, reflecting the robustness requirement.

1.3 YouTube-QA

YouTube-QA [13] comprises 1,970 videos and 50,505 QA pairings extracted from 122,708 natural language descriptions. It covers numerous inquiries, including What, Who, How, Where, When, and Other. Each question corresponds to real-world circumstances documented in YouTube videos, making it well-suited for open-ended and multiple-choice jobs. The dataset includes various difficulties, from essential image identification to advanced reasoning involving motion, interactions, and context. Its extensive covering of themes and real-world focus ensures robust evaluation of VideoQA models in realistic, noisy conditions.

1.4 MovieQA

MovieQA [4] offers 14,944 QA pairs based on movie clips and texts, concentrating on visual-textual story comprehension. Questions frequently need integrating clues from video and text to answer high-level narrative concerns, such as character connections or plot developments. This dataset examines a model’s ability to blend video understanding with script-based textual context. By merging multimodal reasoning with long-form storytelling, MovieQA sets a challenging benchmark for models analyzing deep narrative structures and contextual relationships.

1.5 ActivityNet-QA

ActivityNet-QA [18] comprises longer movies (average 116 seconds) and 58,000 QA pairs focusing on human activities. Questions are divided into Yes/No, Number, Object, Color, and Location. Unique problems include precise spatiotemporal thinking for motion and event understanding. Unlike shorter video datasets, ActivityNet-QA prioritizes understanding sequences of activities and their temporal connections. This dataset is crucial for creating algorithms that evaluate real-world, multi-step actions and reason over time.

1.6 LSMDC-QA

The LSMDC-QA dataset [19] is derived from the Large-Scale Movie Description Challenge and contains 348,998 QA pairs. It specializes in fill-in-the-blank challenges that evaluate a

model's ability to predict missing words or phrases in movie descriptions. Questions are designed to measure temporal dependencies and narrative understanding. As one of the most extensive datasets, it is vital for advancing models capable of analyzing long-term video material and producing contextually appropriate predictions.

1.7 COCO-QA and Visual7W

Initially intended for photos, COCO-QA and Visual7W [20] have been applied to video situations. These datasets focus on basic question categories such as object identification, counting, color recognition, and spatial reasoning. They stress frame-level knowledge, making them essential benchmarks for simpler VideoQA projects. By merging frame-based and question-driven evaluation, they help models that need to generalize static analytic methods to video.

COMPARISON OF THE DATASETS:

Purpose of Dataset Comparison:

2.1 Depth of Scope and Coverage

For example, specific dataset purposes like spatiotemporal reasoning (TGIF-QA) or narrative interpretation (MovieQA). We can now compare datasets to discover which components of VideoQA are taken more seriously: motion understanding, temporal reasoning, or multimodal integration.

2.2 Task Complexity Evaluation

Indeed, the issues varied fairly widely among datasets-from relatively essential item identification to hard state transitions or even fill-in-the-blank narrative tasks. Understanding these distinctions will help pick which benchmark satisfies the purpose of testing specific model competencies.

2.3 Diversity of datasets

Videos in many formats- gifs, short clips, or even long movie form questions can be open-ended, multiple choice, fill-in-the-blank question forms, QA domains: action recognition, state transition, etc. Data Sets Compare datasets to test your models in the most significant possible situations and maximize their robustness.

2.4 Model Performance Benchmarking

Comparisons help bring out the hazards of performance across diverse datasets. A model competent at frame-level tasks, such as COCO-QA, may need to improve temporal reasoning. This understanding helps in the creativity of the development of models tackling various difficulties.

2.5 Relates to Real-world Applications

The best datasets mirror real-world conditions, such as YouTube-QA and Activity Net-QA. Others give controlled task-specific standards like TGIF-QA. Similar datasets exhibit the maximum relevance to real-world applications, guiding researchers toward fascinating

solutions.

Criteria for Dataset Comparison:

Types of Questions

The types of questions included in a dataset dictate its focus and complexity. Questions may be open-ended, multiple-choice, or fill-in-the-blank.

- What: Object or action identification (e.g., "What is the person doing?").
- Who: Actor recognition (e.g., "Who is running?").
- When/Where: It can be either time or space. For example, "When does this happen?"
- How: Analyze motion or action ("How is the person preparing the meal?").
- Why: Evidence-based inquiry that is causal in the nature-for example, "Why is the person angry?

Dataset Examples:

TGIF-QA does involve some activities including repetition counting and state transition recognition.

MovieQA is based on high-level narrative reasoning both in films and texts.

Video Content Characteristics

Short clips are a few seconds, while the lengthy story is a number of minutes.

Example: MSVD-QA employs short-form movies, while ActivityNet-QA includes lengthier footage.

Source: Curated (e.g., TGIF-QA employs animated GIFs).

Real-world (e.g., YouTube-QA contains videos from varied real-world contexts).

Domain: General-purpose datasets (e.g., MSRVTT-QA).

Specific domains, such as movies (e.g., MovieQA) or human activity recognition (e.g., ActivityNet-QA).

Spatio-temporal Complexity

Static vs. Temporal: Static datasets are focused on simple frames or isolated objects. Temporal datasets require reasoning over sequences of frames, such as capturing motion and causation.

Multimodal Integration Modalities

Some datasets are combining written or audio information with video. For example, MovieQA adds video footage to text-based storylines for multimodal inference.
Usefulness: It also encourages building models to incorporate visual, aural, and textual data.

Reasons Behind Dataset Selection:

The VideoQA datasets to compare represent some of the most extensive high-quality benchmarks with each trying to and offering unique chances to test the performance of models in overcoming the most critical difficulties in this domain. Some of the datasets - TGIF-QA, MovieQA, MSVD-QA, and ActivityNet-QA - have widely been adopted by the research community and are often cited in leading academic publications and contests. Their selection ensures that the comparison is relevant and complete, with precise tasks and criteria to compare the disparate models and approaches uniformly. Each of these datasets approaches video from different difficulties in VideoQA, and their combined comparison is necessary to fully

understand the scope of the field. TGIF-QA stresses spatiotemporal reasoning, examining a model's capacity to perceive motion and interactions throughout time. MovieQA, on the other hand, focuses on high-level narrative comprehension, needing models to combine multimodal information from video and textual screenplays. Real-world datasets like YouTube-QA and ActivityNet-QA present chaotic, uncontrolled conditions that replicate practical applications, validating a model's robustness and flexibility in real-world scenarios. By evaluating these datasets together, researchers can get insights into the spectrum of VideoQA issues, from frame-level analysis to sequence-level temporal reasoning and multimodal integration.

The chosen datasets also reflect a wide array of task kinds, ensuring variation in their evaluation scope. For example, TGIF-QA introduces the following new tasks: counting repeated elements and evaluating quantitative reasoning. MovieQA and YouTube-QA not only provide open-ended but also multiple-choice answers, testing the model's capability to process structured and unstructured media. ActivityNet-QA emphasizes long-term temporal and motion reasoning, which is very important for explaining activities distributed over longer videos. Such diversity of tasks will ensure that models are tested along the whole spectrum of their reasoning capacities, from straightforward recognition to more complicated causal and temporal understanding. These datasets additionally differ in maturity and quality. Each dataset has received extreme review and curation towards the credibility of being a benchmark. Datasets like MSVD-QA and MSRVTT-QA, taken from well-established video datasets, provide high-quality annotations and adequate scale for training and evaluation. In contrast, datasets like ActivityNet-QA and MovieQA work with lengthier movies, requiring models to manage temporal connections and high-level reasoning over lengthy time periods. Moreover, datasets like YouTube-QA come with entirely different real-world scenarios, ensuring that the trained and tested models are robust and practical in applications.

These datasets are ideal for comparison also because their applicability goes directly to current approaches. They align strongly with cutting-edge methodologies in VideoQA: spatiotemporal modeling, attention processes, and multimodal learning. Suppose a consideration of these datasets is undertaken. In that case, it allows researchers to critically consider the strengths and limitations of state-of-the-art VideoQA models and pinpoint where innovations are still required.

Lastly, these datasets highlight impressive results of the development of VideoQA research. Novelties such as state transitions and repetition counts in TGIF-QA added a new bar for spatiotemporal thinking. The movie QA established multimodal reasoning by combining video and text narratives, making it the bigger limit from what VideoQA could intend to do. ActivityNet-QA upped the scale of VideoQA, dealing with long-duration videos with real-world applications. Focusing on these datasets, the comparison will not only enlighten the current capacities but, more importantly, highlight the discipline's growth and the prospect of further advancement. It becomes one of the important steps before establishing full knowledge of VideoQA, driving its development ahead.

Table 1: Comparison of mentioned datasets

Dataset	Size	Data Type	QA Tasks	Unique Features	Applications
TGIF-QA	103,919 QA pairs; 56,720 GIFs	Animated GIFs	Repetition Counting, Repeating Action, State Transition, Frame QA	Spatiotemporal reasoning; diverse short tasks	Evaluates motion dynamics and reasoning
YouTube-QA	50,505 QA pairs; 1,970 videos	YouTube videos	Open-ended, multiple-choice	Real-world scenarios, natural language QA	General video comprehension
MSVD-QA	Short-form videos; 50,505 QA pairs	Video clips	What, Who, How, When, where	Short videos; basic QA categories	Object and event recognition
MSRVTT-QA	Long-form videos; extended QA pairs	Video clips	What, Who, How, When, where	Long videos, complex reasoning	Sequential and temporal understanding
ActivityNet-QA	58,000 QA pairs; avg. 116-second videos	Long-form activities	Yes/No, Number, Object, Color, Location	Real-world, multi-step activity reasoning	Human activity comprehension
MovieQA	14,944 QA pairs; movie clips and scripts	Movie clips & text	Open-ended narrative comprehension	Visual-textual story comprehension	Multimodal reasoning, deep narrative tasks
LSMDC-QA	348,998 QA pairs; movie clips	Movie clips	Fill-in-the-blank	Largest dataset; fill-in-the-blank tasks	Temporal dependencies, narrative logic

COCO-QA	Adapted to video contexts	Static images & videos	Object, Number, Colour, Location	Basic QA categories	Frame-level recognition
Visual7W	Adapted to video contexts	Static images & videos	Object, Number, Colour, Location	Spatial reasoning	Object-centric reasoning

EXISTING METHODOLOGIES:

3.1 Dual-LSTM Spatio-Temporal Attention

The Dual-LSTM [11] Spatio-Temporal Attention technique employs a two-layer LSTM network to address VideoQA tasks by encoding video and question-answer pairs. This model contains two fundamental attention mechanisms: spatial attention and temporal attention. Spatial attention focuses on recognizing significant regions inside the video frames, allowing the model to focus on the relevant areas to the stated topic. However, temporal attention favors frames that are most significant in the chronology of the video. The combined features are recovered using ResNet [21] for spatial information and C3D to capture temporal dynamics. These attention mechanisms operate together to match visual and linguistic representations, allowing for successful reasoning over the input video.

This methodology achieved substantial results across numerous tasks, including a mean $\ell 2$ loss of 4.46 for Repetition Counting, 63.79% accuracy for State Transition Identification, 47.79% for Frame QA, and 50.48% for Repeating Action tasks.

The mean ℓ^2 loss is a regression-based metric that measures the squared difference between a model's predicted values and the actual ground truth values. It evaluates how far the predictions deviate from the true values in numerical terms. For example, in the context of VideoQA, where the task involves counting repetitions of an action, the mean $\ell 2$ loss quantifies the error in the predicted count compared to the true count. A loss of 4.46 implies that, on average, the squared difference between the expected and actual counts is 4.46, and the root mean square error ($\sqrt{4.46} \approx 2.11$) represents the typical deviation in the predictions. This model's strength comes from its capacity to integrate spatial and temporal cues simultaneously, yet it has difficulty processing complicated reasoning due to its relatively simplistic architecture.

3.2 Knowledge-Based Progressive Spatial-Temporal Attention Network

K-PSTANet [13] is the recent introduction of both spatial and temporal attention mechanisms with external knowledge sources to improve VideoQA performance. Spatial attention: This module uses a faster R-CNN for object-level feature extraction from the frames, and temporal attention picks those relevant frames to answer queries. Further, the model is infused with a Knowledge Attention mechanism that extracts appropriate external knowledge (DBpedia) to fill in the reasoning gaps in the model. The extracted knowledge is represented using Doc2Vec, which enables the video model to combine information other than that contained in the video footage itself. This progressive architecture refines the joint representation of video, question, and knowledge iteratively while providing very accurate predictions.

The Wu-Palmer Similarity (WUPS) score [22] is a soft generalization of accuracy that allows for ambiguities in anticipated answers. Unlike strict accuracy, which views answers as correct only if they perfectly match the ground truth, WUPS ratings examine the semantic similarity between the expected and actual answers. It utilizes a word-level similarity metric known as the Wu-Palmer similarity, which analyzes the closeness of two concepts in a lexical hierarchy (e.g., WordNet). A WUPS score is produced utilizing criteria (e.g., WUPS@0.9) to measure the significance of a similarity match. For example, the score is punished if the similarity is below the threshold. This makes WUPS particularly useful for instances where approximation answers can still be significant, such as VideoQA.

K-PSTANet scored an overall accuracy of 50.7% (WUPS@0.9 measure) on the YouTube-QA dataset, with task-specific scores of 16.4% for "What" questions, 50.2% for "Who," 79.2% for "How," 50.0% for "Where," and 74.1% for "When." By leveraging external information and question-guided spatiotemporal attention, this methodology tackles weaknesses in purely video-based reasoning and increases the system's capacity to answer complicated queries.

3.3 Two-Stream Spatio-temporal MAC Network

The TS-STMAC [23] is a sophisticated model for advanced spatiotemporal reasoning in VideoQA. It employs a two-stream architecture that helps process video content effectively. The temporal stream retrieves motion-related information from Slow Fast networks that follow the philosophy of tapping dynamic information across clips. The spatial stream utilizes Faster R-CNN for detailed appearance features of objects and areas in every frame. These are then fed into Memory, Attention, and Composition (MAC) cells, which sharpen reasoning over multiple steps iteratively. Each MAC cell [24] links the video content with the question, with step-by-step progressive attention limitation to the most relevant clips and places.

The network additionally combines question-aware attention, employing BERT embeddings and bidirectional LSTMs to encode the question, assuring alignment between visual material and textual queries. Through multi-step reasoning, TS-STMAC excels in addressing complex questions requiring precise temporal and spatial awareness. The model achieved substantial accuracies across many benchmarks: 43.2% on MSVD-QA (best: 33.7% for "What" questions), 39.4% on MSRVTT-QA (best: 78.6% for "When" questions), and 48.3% on ActivityNet-QA. By iteratively enhancing its knowledge, TS-STMAC increases the state of the art in VideoQA, particularly for tasks requiring complicated spatiotemporal reasoning.

Each methodology presents a step-forward approach to addressing the VideoQA task challenges. The 2017 Dual-LSTM Spatio-Temporal Attention provided the base for foundational attention mechanisms through its plain architecture. The 2019 K-PSTANet significantly improved the reasoning by incorporating external knowledge and spatial-temporal attention. Finally, the 2020 TS-STMAC advanced the topic with a two-stream architecture and multi-step reasoning, achieving state-of-the-art results by refining iteratively understanding the temporal and spatial elements. Together, these approaches demonstrate how models are getting more robust at managing increasingly intricate video-related tasks.

SIMILARITIES AMONG THE METHODS:

The commonality in the approaches followed while doing VideoQA. The approaches in the presented VideoQA models share similar ideas and contents, representing the same challenges and requirements of the assignment. While showing differences in design and strategy, all the models converge in some crucial tactics or elements for successfully analyzing video footage and answering the questions.

5.1 Attention Mechanisms

Each relies on attention mechanisms to contend with the richness of the video data. Spatial attention focuses on detecting essential regions inside the frame so that models can pay attention to objects or areas relevant to the query. Temporal attention addresses the sequential nature of films, which aids in focusing on critical time frames. Dual-LSTM Spatio-Temporal Attention attends to both spatial and temporal attention while capturing impressive regions and frames; for K-PSTANet and TS-STMAC, the mechanisms are enhanced with external information and multi-step reasoning.

5.2 Spatial and Temporal Features Integration

Each model requires a combination of spatial and temporal cues. The videos inherently possess spatial (frame-level) and temporal (sequence-level) information. Therefore, this fusion is crucial for thinking. Dual-LSTM injects ResNet for spatial features and C3D for temporal dynamics. Similarly, K-PSTANet and TS-STMAC inject Faster R-CNN for spatial information and the more advanced networks like Slow Fast for motion analysis. That way, the models make sure to evaluate the static as well as dynamic effects efficiently.

5.3 Pretrained Networks

The three approaches leverage the power of earlier pre-trained networks like ResNet, C3D, Faster R-CNN, or Slow Fast for enough feature extraction. These used networks serve as a backbone and encode the visual content leading to powerful representations ready for further processing. In general, depending on the pre-trained models, the approaches reduce the amount of compute burden in training from scratch and also take advantage of the learnt representations of massive datasets.

5.4 Question-Aware Attention

An important constituent in all the methods is question-aware attention. The models connect the video frame retrieved features to the text form of the question. For instance, BERT embeddings and bidirectional LSTMs are utilized for encoding in TS-STMAC. In contrast, K-PSTANet applies external knowledge to improve the context of the video and that of the question.

5.5 Shared Benchmarks for Evaluation

These models are tested on similar benchmarks, including TGIF-QA and MSVD-QA, ensuring uniform performance benchmarking. These datasets test all VideoQA tasks, from simple frame-based queries to complex temporal reasoning, allowing models to show their spatiotemporal understanding.

These methods share some common characteristics: attention mechanisms, integration of spatial

and temporal variables, application of pre-trained networks, question-aware reasoning, and iterative refinement. These tactics explain how state-of-the-art VideoQA systems handle dynamic video content efficiently.

ANALYSIS ON THE METHODOLOGIES:

Comparisons Being Made in the VideoQA Methods. The three methodologies—Dual-LSTM Spatio-Temporal Attention, K-PSTANet, and TS-STMAC —are compared across multiple aspects. These comparisons demonstrate the benefits and limits of each method while emphasizing the advancement of strategies in VideoQA.

6.1 Attention Mechanisms

One of the important comparison areas lies in how these systems use spatial and temporal attention strategies. The Dual-LSTM model applies simple spatial attention to identifying important regions' locations in frames and temporal attention to figuring out important frames. K-PSTANet, on its part, strengthens this by using question-guided attention and external knowledge to dynamically modulate the emphasis on space and time. TS-STMAC takes this further by iteratively applying multi-step attention through Memory, Attention, and Composition (MAC) cells, enabling progressively improved alignment between video content and the inquiry.

6.2 Integration of space and time characteristics

As far as the mentioned methodologies are concerned, the differences in extraction and integration processes can be observed regarding spatial and temporal characteristics derivation. Dual-LSTM uses ResNet to extract spatial information, followed by the acquisition of temporal features using C3D. Such experiments show how sophisticated the models' interpretation of the video content becomes from time to time.

6.3 Reasoning Abilities

Dual-LSTM's simplified architecture allows it to solve basic reasoning tasks but suffers from more complicated queries. K-PSTANet reasons better by using external knowledge to fill the gaps between comprehension, and hence, it can answer a few more advanced queries. TS-STMAC represents the top-most of the three in terms of its reasoning capability as it develops its understanding iteratively through multi-step reasoning. This comparison highlights the move from static reasoning methods to more dynamic and iterative frameworks.

6.4 Use of External Knowledge

One of the major differences is how K-PSTANet uses external sources of knowledge. It covers reasoning gaps that cannot be catered to through video data alone by incorporating knowledge graphs like DBpedia and encoding them with Doc2Vec. In contrast, neither Dual-LSTM nor TS-STMAC uses external knowledge; instead, both rely solely on visual and textual material. This edge allows K-PSTANet to answer better questions involving such external context or semantic understanding.

6.5 Question-aware alignment

Another similarity is how well the models align video aspects to the question. Dual-LSTM

applies basic attention techniques for this alignment, while K-PSTANet incorporates question-guided attention to ensure that the chosen frames and objects align contextually with the question. TS-STMAC uses BERT embeddings as well as bi-directional LSTMs to encode the question, implying even more accurate alignment through iterative multi-step reasoning.

6.6 Evaluation Metrics

Dual-LSTM performs well with simpler metrics like accuracy for both Frame QA and State Transition Identification, whereas for Repetition Counting, which happens to be a regression task, mean ℓ2 loss is used. K-PSTANet incorporates WUPS rankings to consider semantic similarity in responses, which makes it more robust for jobs with approximation or ambiguous responses. TS-STMAC is tested with the help of accuracy metrics on several datasets, such as MSVD-QA and ActivityNet-QA, confirming its capacity to deal with complex tasks. Such differences would indicate the need for task-specific criteria in conducting comparisons over models.

6.7 Coverage of Datasets

The data sets used in the assessment are also a form of comparison. Dual-LSTM has been tested extensively on TGIF-QA, which specializes in short GIFs and spatiotemporal reasoning. K-PSTANet extends the breadth with its application on YouTube-QA, a real dataset with noisy content in videos. TS-STMAC has been evaluated on various datasets, including MSVD-QA, MSRVTT-QA, and ActivityNet-QA, thereby showing its flexibility as well as generalizability. This comparison illustrates how the scope of the datasets being used has been expanding for model evaluations.

LIMITATIONS OF THE METHODOLOGIES:

Despite the success of the three VideoQA approaches Spatio-Temporal Attention, K-PSTANet, and TS-STMAC-there are many limitations that the three approaches share in common. This means much still needs to be addressed when developing robust and highly generalizable VideoQA systems.

7.1 Dependency on Dataset-Specific Features

All these heavily rely on the types of datasets they have been trained and tested with. While Dual-LSTM and K-PSTANet work well on particular feature-contained datasets, such as TGIF-QA or YouTube-QA, they go downhill from there when used on others with very different content. This limitation needs to be fixed across datasets and limits how they can be put to use in real-life applications where there could be varied, unexpected, and unconventional films.

7.2 Restricted Multimodal Integration

For example, K-PSTANet relies on information outside reason, but none fully integrates different information natures: sound, text, and graphics. Many real-world videos contain talking-head-style videos, background noise, and subtitling, none of which these algorithms handle. Not thinking in multiple modes makes it harder for these methods to deal with complicated situations where understanding requires combining information from many sources.

7.3 Troubles in Long-Term Time Understanding

Once more, however, while TS-STMAC embeds complex structures to understand time and

space, all three fail to grasp inner long-range connections in the movie. The datasets, including videos such as ActivityNet-QA, which are long, show how these models cannot learn and maintain context over long times. Their dependency on short-range attention or two-stream structures usually causes a loose or broken understanding of long ranges.

7.4 Computational Complexity

These models incur tremendous computational costs due to their increasing complexity. As such, while TS-STMAC boasts a two-stream architecture with multi-step MAC reasoning, it requires large processing resources for training and inference. Similarly, adding external knowledge in K-PSTANet introduces an extra layer of complexity, making these models less feasible for real-time deployment or resource-constrained contexts.

7.5 Lack of Robustness to Noise

These approaches, especially Dual-LSTM and K-PSTANet, demonstrate diminished performance when applied to noisy or unstructured real-world video data. For example, datasets like YouTube-QA feature videos with changing quality, lighting settings, and background distractions, which might affect model performance. The need for clean, labelled datasets for training renders these methods less effective in addressing a variety of real-world circumstances.

7.6 Mistrusting Open-Ended and Ambiguous Questions

While all three models try to alleviate ambiguity with WUPS scores, K-PSTANet, and its siblings still fail to answer highly open-ended questions when inference is required that is not based on present training data. Thus, the answer to the question "Why is the person smiling?" may involve causal reasoning or background knowledge that is commonly absent from video data and other sources of knowledge. The models are better suited to well-defined factual questions rather than speculative or interpretive ones.

7.7 Limited Interpretability

Although attention mechanisms provide some amount of interpretability, the fundamental workings of these models, particularly in multi-step reasoning frameworks like TS-STMAC, remain opaque. Understanding why a model concentrates on certain frames or regions and how it arrives at its findings is still challenging, making it tougher to trust or debug these systems.

All these recurring challenges call for more generalized, multimodal, and resilient VideoQA models that can handle different real-world scenarios, long-term temporal relationships, and open-ended reasoning tasks while staying computationally efficient and interpretable.

These limitations may be overcome in future research by proposing new methodologies where it can successfully handle the linguistic and vision cues [25,26,27,28] in parallel without much havoc.

Comparison of Methodologies:

Table 2: Comparison of methodologies over accuracy

Methodology	Dataset/Task	Accuracy	WUPS@0.9
Dual-LSTM Spatio-Temporal	Repetition Count (Mean ℓ2)	4.46	N/A
	State Transition	63.79%	N/A
	Frame QA	47.79%	N/A
	Repeating Action	50.48%	N/A
K-PSTANet	Overall (YouTube-QA)	50.7%	50.7%
	What	16.4%	16.4%
	Who	50.2%	50.2%
	How	79.2%	79.2%
	Where	50.0%	50.0%
	When	74.1%	74.1%
TS-STMAC	MSVD-QA	43.2% (Best: "What" 33.7%)	N/A
	MSRVTT-QA	39.4% (Best: "When" 78.6%)	N/A
	ActivityNet-QA	48.3%	N/A

The graphs visually represent the comparative accuracies of the three methodologies—Dual-LSTM Spatio-Temporal Attention, K-PSTANet, and TS-STMAC —across various tasks and datasets.

4.1. Dual-LSTM Spatio-Temporal Attention

Dual-LSTM Spatio-Temporal Attention: The graph depicts the results of the model's performance on TGIF-QA, except for the mean ℓ2 loss metric. It correctly identifies some state transitions (63.79%) as well as frames QA (47.79%) and repeating action (50.48%). Hence, this depicts the model's fundamental capacity to deal with spatiotemporal reasoning but with limitations in complex tasks.

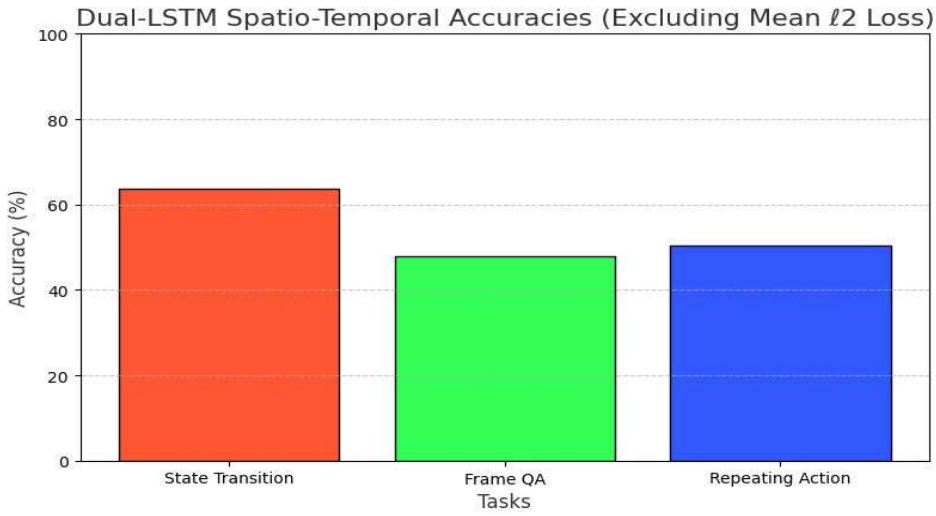

Figure 1: Comparison of different accuracies on Dual-LSTM Spatio-Temporal Attention.

In Figure 1, we compare the accuracies of different tasks performed by the Dual-LSTM and their accuracies.

4.2 Knowledge-Based Progressive Spatial-Temporal Attention Network

K-PSTANet Accuracies (WUPS@0.9): The plot shows the task-specific and overall performance of K-PSTANet on YouTube-QA with its substantial strengths in "How" at 79.2% and "When" at 74.1%. Integrating external knowledge and question-guided attention significantly enhances its reasoning capability compared to simpler architectures.

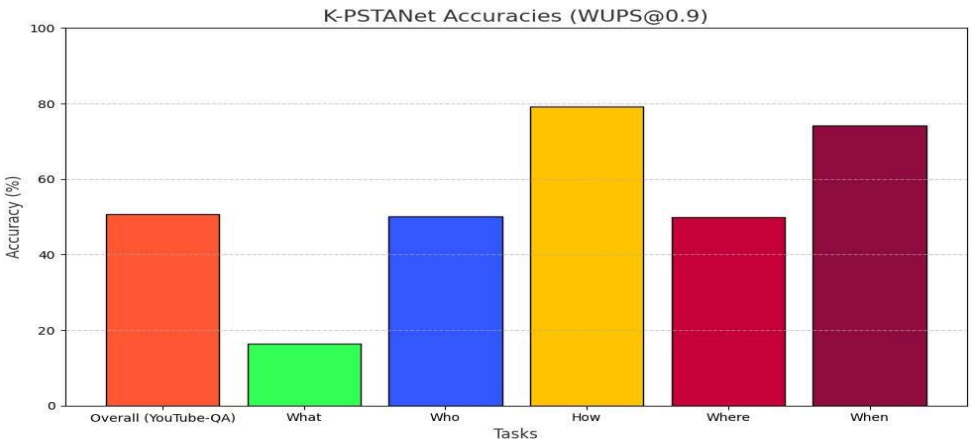

Figure 2: Comparison of different accuracies on K-PSTANet**.**

In Figure 2, we compared the different tasks performed by the K-PSTANet and compared their accuracies.

4.3 Two-Stream Spatiotemporal MAC Network

TS-STMAC Accuracies. TS-STMAC was fairly consistent with the per-dataset accuracies of 43.2%, 39.4%, and 48.3% over MSVD-QA, MSRVTT-QA, and ActivityNet-QA datasets. Its two-stream architecture and multi-step reasoning make it the best approach for complex and diverse datasets and the most generalizable.

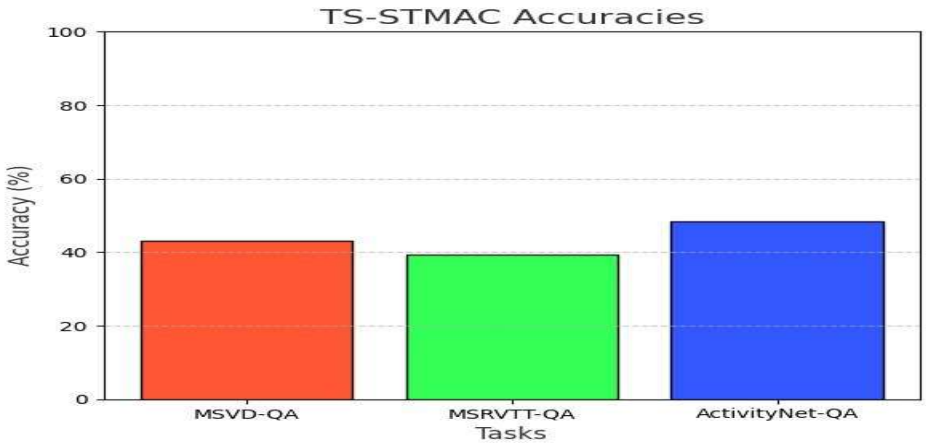

Figure 3. Comparison of different accuracies on TS-STMAC.

In Figure 3, we compared the different Accuracies of Different data sets from the TS-STMAC model.

CONCLUSION:

By way of conclusion, it focuses on three approaches to VideoQA, namely: Dual-LSTM Spatio-

Temporal Attention, K-PSTANet, and TS-STMAC. These show the advancement in spatiotemporal reasoning, attention mechanisms, and integrating external knowledge to handle diverse problems with VideoQA. However, limitations concerning dataset dependence, small multimodal integrations, and difficulties in long-term reasoning underline the necessity of much more robust and generalizable solutions. This analysis reflects the impressive perspective of VideoQA as a tool that links vision and language understanding while pointing to crucial areas for further research and development.

REFERENCES

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar,and F.-F. Li. Large-Scale Video Classification with Convolutional Neural Networks. In CVPR, 2014.

[2] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs.In ICML, 2015.

[3] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, andM. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In ICCV, 2015.

[4] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In CVPR, 2016.

[5] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In EMNLP,2016

[6] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.W. Ha, and B.-T. Zhang. Multimodal Residual Learning for Visual QA. In NIPS, 2016.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In ICML,2015.

[8] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In ICLR,2015.

[9] YunseokJang,YaleSong,YoungjaeYu,YoungjinKim,andGunheeKim.2017.TGIF-QA:Towardspatio-temporalreasoninginvisualquestionanswering.InProceedingsoftheIEEEConferenceonComputerVisionandPatternRecognition(CVPR'17). 1359–1367.

[10]LinchaoZhu,ZhongwenXu,YiYang,andAlexanderG.Hauptmann.2017.Uncoveringthetemporalcontextforvideo question answering.Int.J.Comput.Vis.124,3(2017),409–421.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[12] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. Commun. ACM 57, 10 (2014),78–85.

[13] WEIKEJIN, ZHOU ZHAO, YIMENGLI, JIELI,JUNXIAO, and YUETINGZHUANG. Knowledge-Based

Progressive Spatial-Temporal Attention Network for Video Question Answering. In **ICCV Workshops**, 2019.

[14] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. In NIPS, 2015.

[15] O. Levy and L. Wolf. Live Repetition Counting. In ICCV,2015.

[16] P. Isola, J. J. Lim, and E. H. Adelson. Discovering States and Transformations in Image Collections. In CVPR, 2015.

[17] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang.Video question answering via gradually refined attention over appearance and motion. In Proceedings of the ACM International Conference on Multimedia (ACMMM), pages 1645–1653, 2017.

[18] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pages 9127–9134, 2019.

[19] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal,H. Larochelle, A. Courville, and B. Schiele. Movie Description. IJCV, 2017.

[20] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W:Grounded Question Answering in Images. In CVPR, 2016.

[21]ZhibiaoWuandMarthaPalmer.1994.Verbssemanticsandlexicalselection.InProceedingsofthe32ndAnnualMeeting onAssociationfor Computational Linguistics. Association for ComputationalLinguistics,133–138.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings oftheIEEEConferenceonComputer VisionandPattern Recognition. 770–778.

[23] Taiki Miyanishi, Takuya Maekawa, and Motoaki Kawanabe. Two-Stream Spatiotemporal Compositional Attention Network for VideoQA. In Proceedings of the European Conference on Computer Vision (ECCV), 2020

[24] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[25] Praneel, A. V., & Rao, T. S. (2024). Gated Dual Adaptive Attention Mechanism with Semantic Reasoning, Character Awareness, and Visual-Semantic Ensemble Fusion Decoder for Text Recognition in Natural Scene Images. International Journal of Intelligent Systems and Applications in Engineering, 12(1), 221-234.

[26] Jitendra, M. S.,Shanmuk Srinivas, A. S., Surendra, T., Rao, R. V., & Chowdary, P. R. (2021, October). A study on game development using unity engine. In AIP Conference Proceedings (Vol. 2375, No. 1). AIP Publishing.

[27] Jitendra, M. S., & Radhika, Y. (2023). An ensemble model of CNN with Bi-LSTM for automatic singer identification. Multimedia Tools and Applications, 82(25), 38853-38874.

[28] Kollu, V. V., Amiripalli, S. S., Jitendra, M. S. N. V., & Kumar, T. R. (2021). A network science-based performance improvement model for the airline industry using NetworkX. International Journal of Sensors Wireless Communications and Control, 11(7), 768-773.